

# Multivariate analysis of mixed data:

## The PCAmixdata R package

Marie Chavent<sup>1,2</sup>, Vanessa Kuentz-Simonet<sup>3</sup>, Amaury Labenne<sup>3</sup>, Jérôme Saracco<sup>2,4</sup>

December 5, 2014

<sup>1</sup> Université de Bordeaux, IMB, CNRS, UMR 5251, France

<sup>2</sup> INRIA Bordeaux Sud-Ouest, CQFD team, France

<sup>3</sup> Irstea, UR ETBX, France

<sup>4</sup> Institut Polytechnique de Bordeaux, France

### Abstract

Mixed data type arise when observations are described by a mixture of numerical and categorical variables. The R package PCAmixdata extends standard multivariate analysis methods to incorporate this type of data. The key techniques included in the package are PCAmix (PCA of a mixture of numerical and categorical variables), PCArot (rotation in PCAmix) and MFAmix (multiple factor analysis with mixed data within a dataset). This paper gives a synthetic presentation of the three algorithms with details and elements of proof to help the user to well understand graphical and numerical outputs of the package. The three main procedures are illustrated on real data composed of four datasets characterizing conditions of life of cities of Gironde, a south-west region of France.

**Keywords:** mixture of numerical and categorical variables, principal component analysis, multiple correspondence analysis, multiple factor analysis, rotation.

## 1 Introduction

Multivariate data analysis refers to descriptive statistical methods used to analyze data arising from more than one variable. These variables can be either numerical or categorical. For example, principal component analysis (PCA) handles numerical variables whereas multiple correspondence analysis (MCA) handles categorical variables. Multiple factor analysis (MFA; Escofier and Pagès,

1994; Abdi et al., 2013) works with multi-table data where the type of the variables can vary from one dataset to the other but the variables should be of the same type within a given dataset. Several existing R packages implement standard multivariate analysis methods. These include Ade4 (Dray et al., 2007), FactoMineR (Lê et al., 2008) or ExPosition (Beaton et al., 2014). However, none of these applications are dedicated to multivariate analysis of mixed data.

Our new R package, PCAmixdata, contains extensions of standard multivariate analysis methods for mixed data. It proposes a number of core techniques: PCAmix (PCA of a mixture of numerical and categorical variables), PCArrot (rotation in PCAmix), MFAmix (multiple factor analysis with mixed data within a dataset). The PCAmix and PCArrot procedures were first proposed in Chavent et al. (2012). In this paper, the presentation of PCAmix is based on generalized singular value decomposition (GSVD) with new mathematical justifications. In addition, the presentation of PCArrot is simplified. The MFAmix procedure handles a mixture of numerical and categorical variables within a group - something which was not possible in the standard MFA procedure. This paper provides a synthetic presentation of these three algorithms, including details and elements of proof to help the user to understand the graphical and numerical outputs of the package. We also specify techniques to project new observations onto the principal components of the three methods. Moreover, the methods and package are illustrated on a real dataset composed of four datatables, each characterizing living conditions in  $n = 542$  cities in the Gironde region in southwest France. This dataset, available in the R package PCAmixdata, was taken from the 2009 census database<sup>1</sup> of the French national institute of statistics and economic studies and from a topographic database<sup>2</sup> of the French national institute of geographic and forestry information. The first datatable describes 542 cities in the Gironde region with 9 numerical variables relating to employment conditions. The second datatable describes those cities with 5 variables (2 categorical and 3 numerical) relating to housing conditions, the third one with 9 categorical variables relating to services (restaurants, doctors, post offices,...) and the last one with 4 numerical variables relating to environmental conditions. A complete description of the 27 variables, divided into 4 groups (Employment, Housing, Services, Environment) is given in Appendix B.

The rest of the paper is organized as follows. Section 2 comprises a detailed examination of the link between GSVD, PCA and MCA. It demonstrates how MCA can be obtained from a single PCA, the cornerstone for merging standard PCA and MCA in PCAmix. Sections 3, 4 and 5 present (respectively) the PCAmix, PCArrot and MFAmix methods. For each procedure a presentation of the algorithm, a description of the numerical and graphical outputs, and an illustration on the real

---

<sup>1</sup><http://www.insee.fr/fr/bases-de-donnees/>

<sup>2</sup><http://professionnels.ign.fr/bdtopo>

data are all provided. Mathematical proof, as well as the R code referred to in this paper, can be found in the appendices.

## 2 GSVD and PCA with metrics

The PCA with metrics is a generalization of the standard PCA method. It is based on the GSVD of a real data matrix  $\mathbf{Z}$ . Metrics are used to introduce weights to rows and columns of  $\mathbf{Z}$  in PCA. In this paper, standard PCA and MCA are presented within this framework, so that a unique PCA procedure can be defined for numerical and categorical data, including the appropriate metrics.

### 2.1 GSVD of real matrix $\mathbf{Z}$

GSVD provides a matrix decomposition of  $\mathbf{Z}$  of dimension  $n \times p$  using two positive definite square matrices  $\mathbf{N}$  and  $\mathbf{M}$ , where  $\mathbf{N}$  is a metric on  $\mathbb{R}^n$  and  $\mathbf{M}$  is a metric on  $\mathbb{R}^p$ . The GSVD of  $\mathbf{Z}$  with metrics  $\mathbf{N}$  and  $\mathbf{M}$  gives the following decomposition:

$$\mathbf{Z} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^t, \quad (1)$$

where

- $\mathbf{\Lambda} = \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_r})$  is the  $r \times r$  diagonal matrix of the singular values of  $\mathbf{Z}\mathbf{M}\mathbf{Z}^t\mathbf{N}$  and  $\mathbf{Z}^t\mathbf{N}\mathbf{Z}\mathbf{M}$ , and  $r$  denotes the rank of  $\mathbf{Z}$ ;
- $\mathbf{U}$  is the  $n \times r$  matrix of the first  $r$  eigenvectors of  $\mathbf{Z}\mathbf{M}\mathbf{Z}^t\mathbf{N}$  such that  $\mathbf{U}^t\mathbf{N}\mathbf{U} = \mathbb{I}_r$ , with  $\mathbb{I}_r$  the identity matrix of size  $r$ ;
- $\mathbf{V}$  is the  $p \times r$  matrix of the first  $r$  eigenvectors of  $\mathbf{Z}^t\mathbf{N}\mathbf{Z}\mathbf{M}$  such that  $\mathbf{V}^t\mathbf{M}\mathbf{V} = \mathbb{I}_r$ .

Note that the GSVD of  $\mathbf{Z}$  can be obtained by performing the standard SVD of the matrix  $\tilde{\mathbf{Z}} = \mathbf{N}^{1/2}\mathbf{Z}\mathbf{M}^{1/2}$ , that is a GSVD with metrics  $\mathbb{I}_n$  on  $\mathbb{R}^n$  and  $\mathbb{I}_p$  on  $\mathbb{R}^p$ . It gives:

$$\tilde{\mathbf{Z}} = \tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{V}}^t \quad (2)$$

and transformation back to the original scale gives:

$$\mathbf{\Lambda} = \tilde{\mathbf{\Lambda}}, \quad \mathbf{U} = \mathbf{N}^{-1/2}\tilde{\mathbf{U}}, \quad \mathbf{V} = \mathbf{M}^{-1/2}\tilde{\mathbf{V}}. \quad (3)$$

### 2.2 PCA of $\mathbf{Z}$ with metrics

GSVD can be used to introduce weights to rows and columns of  $\mathbf{Z}$  in PCA. The associated metrics  $\mathbf{N}$  and  $\mathbf{M}$  are the diagonal matrices of those weights. The factor scores of the rows and the factor scores of the columns are then obtained as follows.

**Factor scores of the rows.** Let  $\mathbf{F}$  denote the  $n \times r$  factor scores matrix of the rows. The scores are the coordinates of the orthogonal projections with respect to the inner product matrix  $\mathbf{M}$  of the  $n$  rows of  $\mathbf{Z}$  onto the axes spanned by the vectors  $\mathbf{v}_1, \dots, \mathbf{v}_r$  (columns of  $\mathbf{V}$ ). This definition gives:

$$\mathbf{F} = \mathbf{Z}\mathbf{M}\mathbf{V}. \quad (4)$$

We deduce from (1) that:

$$\mathbf{F} = \mathbf{U}\mathbf{\Lambda}. \quad (5)$$

Recall that the columns of  $\mathbf{V}$  are the eigenvectors of  $\mathbf{Z}^t\mathbf{N}\mathbf{Z}\mathbf{M}$  which can be found by solving the sequence (indexed by  $i$ ) of optimization problems:

$$\begin{aligned} & \text{maximize} && \|\mathbf{Z}\mathbf{M}\mathbf{v}_i\|_{\mathbf{N}}^2 \\ & \text{subject to} && \mathbf{v}_i^t\mathbf{M}\mathbf{v}_j = 0 \quad \forall 1 \leq j < i, \\ & && \mathbf{v}_i^t\mathbf{M}\mathbf{v}_i = 1. \end{aligned} \quad (6)$$

Here  $\|\mathbf{x}\|_{\mathbf{N}}^2 = \mathbf{x}^t\mathbf{N}\mathbf{x}$ . Let us denote  $\mathbf{f}_i = \mathbf{Z}\mathbf{M}\mathbf{v}_i$  a column of  $\mathbf{F}$ . The vectors  $\mathbf{v}_1, \dots, \mathbf{v}_r$  are then defined in such a way that  $\|\mathbf{f}_i\|_{\mathbf{N}}^2 = \lambda_i$  is maximum. The columns of  $\mathbf{F}$  are called the principal components (PC).

**Factor scores of the columns.** Let  $\mathbf{A}$  denote the  $p \times r$  factor scores matrix of the columns. The scores are the coordinates of the orthogonal projections with respect to the inner product matrix  $\mathbf{N}$  of the  $p$  columns onto the axes spanned by the vectors  $\mathbf{u}^1, \dots, \mathbf{u}^r$  (columns of  $\mathbf{U}$ ). This definition gives:

$$\mathbf{A} = \mathbf{Z}^t\mathbf{N}\mathbf{U}. \quad (7)$$

We deduce from (1) that:

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}. \quad (8)$$

Recall that the columns of  $\mathbf{U}$  are the eigenvectors of  $\mathbf{Z}\mathbf{M}\mathbf{Z}^t\mathbf{N}$  which can be found by solving the sequence (indexed by  $i$ ) of optimization problems:

$$\begin{aligned} & \text{maximize} && \|\mathbf{Z}^t\mathbf{N}\mathbf{u}_i\|_{\mathbf{M}}^2 \\ & \text{subject to} && \mathbf{u}_i^t\mathbf{N}\mathbf{u}_j = 0 \quad \forall 1 \leq j < i, \\ & && \mathbf{u}_i^t\mathbf{N}\mathbf{u}_i = 1. \end{aligned} \quad (9)$$

Let us denote  $\mathbf{a}_i = \mathbf{Z}^t\mathbf{N}\mathbf{u}_i$  a column of  $\mathbf{A}$ . The vectors  $\mathbf{u}_1, \dots, \mathbf{u}_r$  are then defined in such a way that  $\|\mathbf{a}_i\|_{\mathbf{M}}^2 = \lambda_i$  is maximum. The columns of  $\mathbf{A}$  are also called the loadings.

The following result will be useful for the orthogonal rotation technique described in Section 4. Note that  $\tilde{\mathbf{A}} = \mathbf{A}$  in the standard SVD decomposition of  $\tilde{\mathbf{Z}} = \mathbf{N}^{1/2}\mathbf{Z}\mathbf{M}^{1/2}$  in (2). It gives:

$$\lambda_i = \|\mathbf{a}_i\|_{\mathbf{M}}^2 = \|\tilde{\mathbf{a}}_i\|_{\mathbb{I}_p}^2$$

where  $\tilde{\mathbf{a}}_i$  is the  $i$ th column of  $\tilde{\mathbf{A}} = \tilde{\mathbf{V}}\tilde{\mathbf{A}}$ .

### 2.3 Standard PCA and MCA

This section presents how standard PCA (for numerical data) and standard MCA (for categorical data) can be obtained from the GSVD of specific matrices  $\mathbf{Z}$ ,  $\mathbf{N}$ ,  $\mathbf{M}$ . The numerical matrix  $\mathbf{Z}$  is obtained by pre-processing of the original data matrix and the matrix  $\mathbf{N}$  (resp.  $\mathbf{M}$ ) is the diagonal matrix of the weights of the rows (resp. the columns) of  $\mathbf{Z}$ .

**Standard PCA.** The data table to be analyzed by PCA comprises  $n$  observations described by  $p$  numerical variables and it is represented by the  $n \times p$  matrix  $\mathbf{X}$ . In the pre-processing step, the columns of  $\mathbf{X}$  are centered and normalized to construct the standardized matrix  $\mathbf{Z}$  (defined such that  $\frac{1}{n}\mathbf{Z}^t\mathbf{Z}$  is the linear correlation matrix). The  $n$  rows (observations) are weighted by  $\frac{1}{n}$  and the  $p$  columns (variables) are weighted by 1. It gives  $\mathbf{N} = \frac{1}{n}\mathbb{I}_n$  and  $\mathbf{M} = \mathbb{I}_p$ . The metric  $\mathbf{M}$  indicates that the distance between two observations is the standard euclidean distance between two rows of  $\mathbf{Z}$ . The total inertia of  $\mathbf{Z}$  is then equal to  $p$ . The factor scores matrix  $\mathbf{F}$  (scores of the observations) and the factor scores matrix  $\mathbf{A}$  (scores of the variables) are calculated directly from (5) and (8). The well-known properties of PCA are the following.

- Each score  $a_{ji}$  is the linear correlation between the numerical variable  $\mathbf{x}_j$  (the  $j$ th column of  $\mathbf{X}$ ) and the  $i$ th principal component  $\mathbf{f}_i$  (the  $i$ th column of  $\mathbf{F}$ ):

$$a_{ji} = \mathbf{z}_j^t \mathbf{N} \mathbf{u}_i = r(\mathbf{x}_j, \mathbf{f}_i), \quad (10)$$

with  $\mathbf{u}_i = \frac{\mathbf{f}_i}{\lambda_i}$  and  $\mathbf{z}_j$  (resp.  $\mathbf{x}_j$ ) the  $j$ th column of  $\mathbf{Z}$  (resp.  $\mathbf{X}$ ).

- Each eigenvalue  $\lambda_i$  is the variance of the  $i$ th principal component:

$$\lambda_i = \|\mathbf{f}_i\|_{\mathbf{N}}^2 = \text{Var}(\mathbf{f}_i). \quad (11)$$

- Each eigenvalue  $\lambda_i$  is also the sum of the squared correlations between the  $p$  variables and the  $i$ th principal component:

$$\lambda_i = \|\mathbf{a}_i\|_{\mathbf{M}}^2 = \sum_{j=1}^p r^2(\mathbf{x}_j, \mathbf{f}_i). \quad (12)$$

**Standard MCA.** The data table to be analyzed by MCA comprises  $n$  observations described by  $p$  categorical variables and it is represented by the  $n \times p$  matrix  $\mathbf{X}$ . Each categorical variable has  $m_j$  levels and the sum of the  $m_j$  is equal to  $m$ . In the pre-processing step each level is coded as a binary variable and the  $n \times m$  indicator matrix  $\mathbf{G}$  is constructed. Usually, MCA is performed by applying standard Correspondence Analysis (CA) to the indicator matrix. In CA the factor scores of the rows and (respectively) columns of the data table are obtained by applying PCA on two different matrices: the matrix of the row profiles and the matrix of the column profiles. Here, we provide a way to calculate the factor scores of MCA from a single PCA with metrics.

Let  $\mathbf{Z}$  now denote the centered indicator matrix. The  $n$  rows (observations) are weighted by  $\frac{1}{n}$  and the  $m$  columns (levels) are weighted by  $\frac{n}{n_s}$ , the inverse of the frequency of the level  $s$ , where  $n_s$  denotes the number of observations that belong to the  $s$ th level. It gives  $\mathbf{N} = \frac{1}{n}\mathbb{I}_n$  and  $\mathbf{M} = \text{diag}(\frac{n}{n_s}, s = 1 \dots, m)$ . This metric  $\mathbf{M}$  indicates that the distance between two observations is a weighted euclidean distance in the spirit of the  $\chi^2$  distance in CA. This distance gives more importance to rare levels. The total inertia of  $\mathbf{Z}$  with this distance and the weights  $\frac{1}{n}$  is equal to  $m - p$ . The GSVD of  $\mathbf{Z}$  with these metrics allows to calculate directly the matrix  $\mathbf{F}$  of the factor scores of the observations from (5). The factor scores of the levels however are not calculated directly from (8). Let  $\mathbf{A}^*$  denote the matrix of the factor scores of the levels. We have:

$$\mathbf{A}^* = \mathbf{M}\mathbf{V}\mathbf{\Lambda}. \quad (13)$$

Note that this result is different from the result of PCA with metrics where the scores of the columns are given by  $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}$ . The proof of this result is given in Appendix A.

The usual properties of MCA are the following.

- Each score  $a_{si}^*$  is the mean value of the (normalized) factor scores of the observations that belong to level  $s$  :

$$a_{si}^* = \frac{n}{n_s} a_{si} = \frac{n}{n_s} \mathbf{z}_s^t \mathbf{N} \mathbf{u}_i = \bar{u}_i^s, \quad (14)$$

with  $\mathbf{z}_s$  the  $s$ th column of  $\mathbf{Z}$  and  $\bar{u}_i^s$  the mean value of the components of  $\mathbf{u}_i$  associated with the observations that belong to level  $s$ .

- Each eigenvalue  $\lambda_i$  is the sum of the correlation ratios between the  $p$  categorical variables and the  $i$ th principal component (which is numerical):

$$\lambda_i = \|\mathbf{a}_i\|_{\mathbf{M}}^2 = \|\mathbf{a}_i^*\|_{\mathbf{M}^{-1}}^2 = \sum_{j=1}^p \eta^2(\mathbf{f}_i | x_j). \quad (15)$$

The correlation ratio  $\eta^2(\mathbf{f}_i | x_j)$  measures the part of the variance of  $\mathbf{f}_i$  explained by the categorical variable  $j$ .

This way, compared to standard MCA calculated by applying CA to the indicator matrix, we can notice that:

- the total inertia is multiplied by  $p$  and is equal to  $m - p$ . This will be useful in PCA for mixed data to balance the inertia of the numerical data (equal to the number of numerical variables) and the inertia of the categorical data (equal now to the number of levels minus the number of categorical variables).
- the scores of the levels are the same. However, the eigenvalues are multiplied by  $p$  and observation scores are then multiplied by  $\sqrt{p}$ . This property has no impact on interpretation since results are identical to within one multiplier coefficient.

### 3 PCA of mixed data

Several methods for PCA of mixed data already exist. For example, the `dudi.mix` function of the ADE4 R package implements the method developed by Hill and Smith (1976). Another example is the `AFDM` function of the FactoMineR R package, which uses the method designed by Pagès (2004). However, the PCAmix method proposed in this paper differs from the aforementioned approaches in that it uses a generalized singular value decomposition (GSVD) of preprocessed data. This GSVD approach includes standard PCA and MCA as special cases.

#### 3.1 The PCAmix method

The data table to be analyzed by PCAmix comprises  $n$  observations described by  $p_1$  numerical variables and  $p_2$  categorical variables. It is represented by the  $n \times p_1$  numerical matrix  $\mathbf{X}_1$  and the  $n \times p_2$  categorical matrix  $\mathbf{X}_2$ . Let  $m$  denote the total number of levels of the  $p_2$  categorical variables. The PCAmix method is a two steps procedure, involving the merging of standard PCA and MCA as described below.

##### Step 1: the pre-processing phase.

1. Build the real matrix  $\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2]$  of dimension  $n \times (p_1 + m)$  where:
  - $\hookrightarrow \mathbf{Z}_1$  is the standardized version of  $\mathbf{X}_1$  (as in PCA),
  - $\hookrightarrow \mathbf{Z}_2$  is the centered version of the indicator matrix  $\mathbf{G}$  of  $\mathbf{X}_2$  (as in standard MCA).
2. Build the diagonal matrix  $\mathbf{N}$  of the weights of the rows of  $\mathbf{Z}$ . The  $n$  rows are weighted by  $\frac{1}{n}$ , such that  $\mathbf{N} = \frac{1}{n} \mathbb{I}_n$ .

3. Build the diagonal matrix  $\mathbf{M}$  of the weights of the columns of  $\mathbf{Z}$ :

$\hookrightarrow$  The  $p_1$  first columns are weighted by 1 (as in PCA).

$\hookrightarrow$  The  $m$  last columns are weighted by  $\frac{n}{n_s}$  (as in MCA), where  $n_s, s = 1, \dots, m$  denotes the number of observations that belong to the  $s$ th level.

This metric  $\mathbf{M} = \text{diag}(1, \dots, 1, \frac{n}{n_1}, \dots, \frac{n}{n_m})$  indicates that the distance between two rows of  $\mathbf{Z}$  is a mixture of the simple euclidean distance used in PCA (for the first  $p_1$  columns) and the weighted distance in the spirit of the  $\chi^2$  distance used in MCA (for the last  $m$  columns). The total inertia of  $\mathbf{Z}$  with this distance and the weights  $\frac{1}{n}$  is then equal to  $p_1 + m - p_2$ .

**Step 2: the factor scores processing step.**

1. The GSVD of  $\mathbf{Z}$  with metrics  $\mathbf{N}$  and  $\mathbf{M}$  gives the decomposition:

$$\mathbf{Z} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^t$$

as defined in Subsection 2.1.

2. The set of factor scores for rows ( $n$  observations) is defined as:

$$\mathbf{F} = \mathbf{Z}\mathbf{M}\mathbf{V}, \tag{16}$$

or directly computed from the GSVD decomposition as:

$$\mathbf{F} = \mathbf{U}\mathbf{\Lambda}. \tag{17}$$

3. The set of factor scores for columns is computed as:

$$\mathbf{A}^* = \mathbf{M}\mathbf{V}\mathbf{\Lambda}. \tag{18}$$

The matrix  $\mathbf{A}^*$  is divided up as follows:  $\mathbf{A}^* = \begin{bmatrix} \mathbf{A}_1^* \\ \mathbf{A}_2^* \end{bmatrix} \begin{matrix} \} p_1 \\ \} m \end{matrix}$  where

$\hookrightarrow \mathbf{A}_1^*$  contains the factor scores of the  $p_1$  numerical variables,

$\hookrightarrow \mathbf{A}_2^*$  contains the factor scores of the  $m$  levels.

Note that in case of purely numerical data, PCAmix is a standard PCA, whereas in the case of only categorical data, it is a MCA. Where there is a mixture of numerical and categorical data, the properties used for the interpretation of graphical outputs in PCA and in MCA remain correct.



### 3.2 Graphical outputs of PCAmix

PCAmix (as PCA and MCA) represents the pattern of similarity of observations and variables by displaying them as points on maps. Observations, numerical variables and levels can be plotted as points in the component space using their factor scores as coordinates. The properties used in PCA or in MCA to interpret these component maps (see section 2.3) are also applicable to PCAmix.

- The factor scores of the  $p_1$  numerical variables (the  $p_1$  first rows of  $\mathbf{A}^*$ ) are correlations with the principal components (the columns of  $\mathbf{F}$ ) as in PCA.
- The factor scores of the  $m$  levels (the  $m$  last rows of  $\mathbf{A}^*$ ) are mean values of the (normalized) factor scores of the observations that belong to these levels as in MCA.

These two properties are important in interpreting the component map of the observations according to the component map of the numerical variables, as well as that of the levels of the categorical variables.

Hereafter, we introduce the notion of “squared loading” which is an important part of the outputs of the three methods implemented in the package. The use of this term is justified by the need for categorical variables to draw an analogy with the notion of loading in PCA. The term “squared loading” is generic for numerical and categorical variables. It provides a way of measuring the link between variables (regardless of type) and principal components. If the variable  $j$  is numerical the squared loading is the squared correlation  $r^2(\mathbf{f}_i, \mathbf{x}_j)$ . If this same variable is categorical the squared loading is the correlation ratio  $\eta^2(\mathbf{f}_i | \mathbf{x}_j)$ . These two measures vary between 0 and 1 and give an idea of the link between the variable and the principal component. These values are represented in the graphical outputs of PCAmix, PCArot and MFAmix for the simultaneous plot of the numerical and categorical variables.

For PCAmix, the “squared loading” of a variable is equal to its contribution to the component considered. Indeed the contribution of a variable to a component is the part of the variance of this component explained by the variable. It was shown in Section 2.2 that  $\lambda_i = \|\mathbf{f}_i\|_{\mathbf{N}}^2 = \text{Var}(\mathbf{f}_i)$ . Moreover as  $\lambda_i = \|\mathbf{a}_i\|_{\mathbf{M}}^2$ , the contribution can therefore be calculated directly from the matrix  $\mathbf{A}$  (instead of  $\mathbf{A}^*$ ). Let  $c_{ji}$  denote the contribution of the variable  $j$  to the component  $i$ . We have:

$$\begin{cases} c_{ji} = a_{ji}^2 = a_{ji}^{*2} & \text{if the variable } j \text{ is numerical,} \\ c_{ji} = \sum_{s \in I_j} \frac{n_s}{n} a_{si}^2 = \sum_{s \in I_j} \frac{n_s}{n} a_{si}^{*2} & \text{if the variable } j \text{ is categorical,} \end{cases} \quad (19)$$

where  $I_j$  is the set of indices of the levels of the variable  $j$ .

### 3.3 Numerical outputs of PCAmix

PCAmix computes new numerical variables called principal components that will “explain” or “extract” the largest part of the inertia of the data table  $\mathbf{Z}$ . The principal components (columns of  $\mathbf{F}$ ) are then non correlated linear combinations of the columns of  $\mathbf{Z}$  and can be viewed as new synthetic variables with:

- maximum dispersion:  $\lambda_i = \|\mathbf{f}_i\|_{\mathbf{N}}^2 = \text{Var}(\mathbf{f}_i)$ ,
- maximum link to the original variables:

$$\lambda_i = \|\mathbf{a}_i\|_{\mathbf{M}}^2 = \sum_{j=1}^{p_1} r^2(\mathbf{f}_i, \mathbf{x}_j) + \sum_{j=p_1+1}^{p_2} \eta^2(\mathbf{f}_i | \mathbf{x}_j). \quad (20)$$

**Prediction of the scores of new observations.** The coefficients of these linear combinations are useful in projecting new observations onto the principal components of PCAmix. According to the definition of the metric  $\mathbf{M}$  given in Subsection 3.1, the linear combination of the vectors  $\mathbf{z}_1, \dots, \mathbf{z}_{p_1+m}$  (columns of  $\mathbf{Z}$ ) for the  $i$ th principal component is:

$$\mathbf{f}_i = \mathbf{ZMv}_i = \sum_{k=1}^{p_1} v_{ki} \mathbf{z}_k + \sum_{k=p_1+1}^{p_1+m} \frac{n}{n_k} v_{ki} \mathbf{z}_k.$$

It is easy to show that  $\mathbf{f}_i$  writes:

$$\mathbf{f}_i = \beta_0 + \sum_{k=1}^{p_1+m} \beta_k \mathbf{x}_k$$

where the vectors  $\mathbf{x}_1, \dots, \mathbf{x}_{p_1+m}$  are the columns of  $\mathbf{X} = (\mathbf{X}_1 | \mathbf{G})$ . It gives:

$$\begin{aligned} \beta_0 &= - \sum_{l=1}^{p_1} v_{li} \frac{\bar{\mathbf{x}}_l}{\sigma_l} - \sum_{l=p_1+1}^{p_1+m} v_{li} \frac{n}{n_l} \bar{\mathbf{x}}_l, \\ \beta_k &= v_{ki} \frac{1}{\sigma_k}, \text{ for } k = 1, \dots, p_1, \\ \beta_k &= v_{ki} \frac{n}{n_k}, \text{ for } k = p_1 + 1, \dots, p_1 + m, \end{aligned}$$

with  $\bar{\mathbf{x}}_k$  and  $\sigma_k$  respectively denote the empirical mean and standard deviation of the column  $\mathbf{x}_k$ .

### 3.4 Illustration of PCAmix

The procedure PCAmix is illustrated with the R package PCAmixdata using the real data **gironde** made up of four datasets. The dataset **housing** describes cities in the Gironde region with categorical and numerical variables on housing conditions. The function **PCAmix** calculates the factor

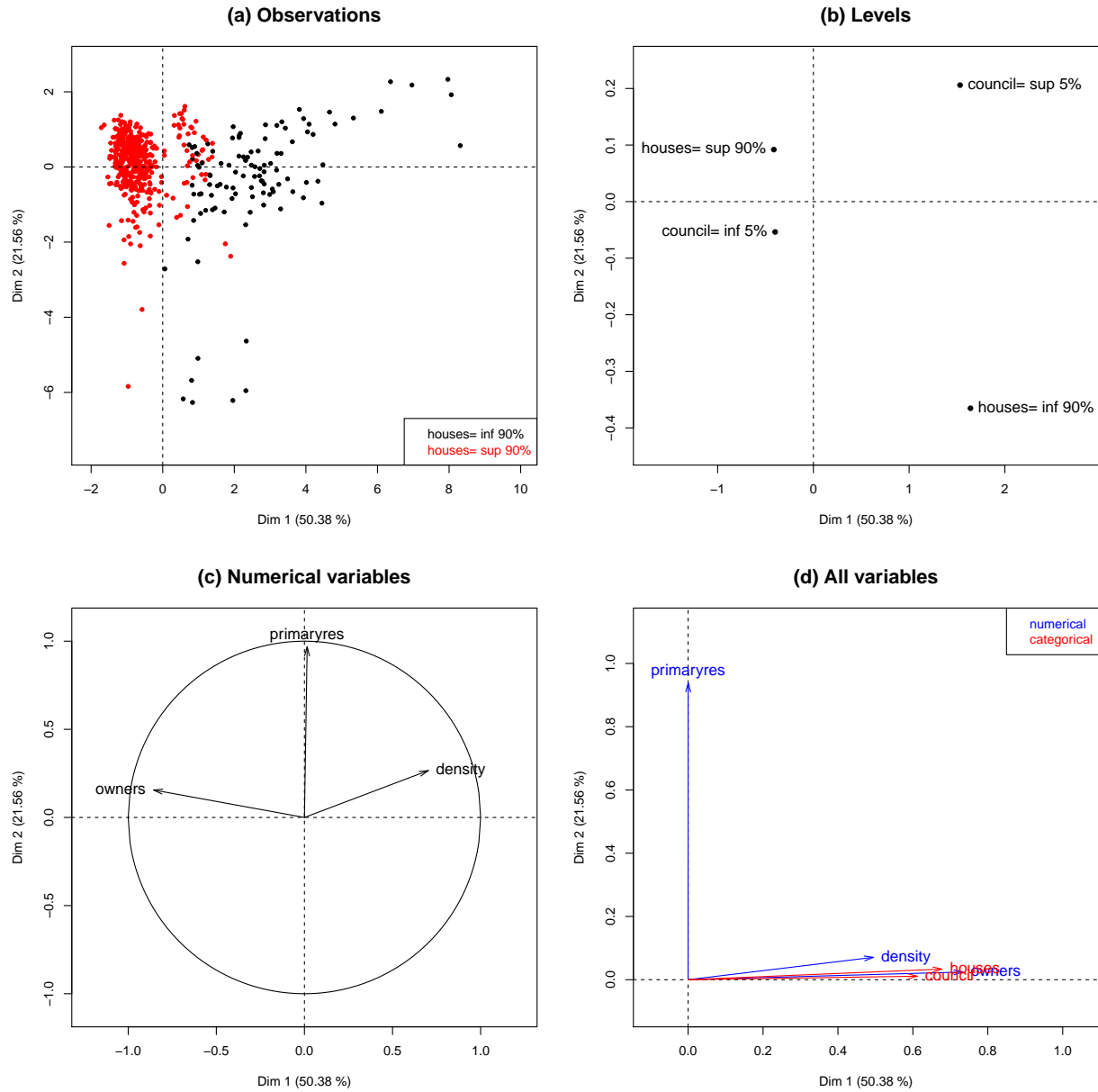


Figure 1: Graphical outputs of PCAmix for the dataset housing. (a) Component map with factor scores of cities. (b) Component map with factor scores of levels. (c) Component map with factor scores of numerical variables. (d) Plot of the squared loadings of all variables.

scores of the 542 cities, the factor scores of the 4 levels of the two categorical variables and the factor scores of the 3 numerical variables. The function `plot.PCAmix` provides a number of graphical outputs. The complete R code required to run this example is provided in Appendix C.

Fig. 1(a) shows a component map of the rows (cities) which are color coded according to their percentage of houses (less than 90%, more than 90%). The first dimension highlights cities with high proportions of houses on the left. Fig. 1(b) shows a component map of the categorical columns (levels), which suggests that cities with a high proportion of houses (on the left) have a low percentage of council housing. The component map of numerical variables shown in Fig. 1(c), referred to as the circle of correlation, indicates that population density is negatively correlated with the percentage of home owners and that these two variables discriminate the cities on the first dimension. Fig. 1(d) is a plot of all the variables (categorical or numerical) in which squared loadings are used as coordinates. For numerical variables, squared loadings are squared correlations and for categorical variables the squared loadings are correlation ratios. In both cases, they measure the link between variables and principal components. In PCAmix, squared loadings are also here absolute contributions to the variance of the components.

In this illustration, only the first two components can be interpreted. The third does not provide good quality projection of the variables. One observes that the three variables “houses”, “council” and “owners” are linked to each other and to the first component. On the contrary, the variable of the percentage of primary residence is clearly orthogonal to these variables and associated to the second component. The variable “density” is not well projected onto this first factorial plane (the squared cosine on the first component is equal to 0.5). These outputs show that the first component will separate cities based on type of housing. On the right of the first factorial plane, cities have a relatively high proportion of council housing and a small percentage of houses. In this case, housing is made up mainly of rented accommodation. On the contrary the type of housing of cities on the left is mostly composed of owners of their houses. The percentage of primary residences has also a structuring role in the characterization of cities in this region of France, by defining towns based on the type of housing to be found within them-either main residences or second homes. Note that a function `predict.PCAmix` is also available in the package to calculate factor scores of new cities not used in the PCAmix procedure.

## 4 Orthogonal rotation in PCA of mixed data

The principal component analysis extracts a set of factors from the dataset. In general, only a subset of  $k$  factors is kept for further consideration. In order to simplify the interpretation of these

$k$  factors, a rotation procedure can be applied. Varimax rotation (Kaiser, 1958) is a very popular orthogonal rotation method for numerical data (that is for standard PCA). A simple solution means here that each factor has a small number of large loadings and a large number of zero (or small) loadings. This simplifies the interpretation because, after a varimax rotation, each original variable tends to be associated with one (or a small number) of factors, and each factor represents only a small number of variables.

For mixed data type (that is for PCAmix), Chavent et al. (2012) proposed an efficient “varimax type” orthogonal rotation algorithm, based on an approach first developed by Kiers (1991). The varimax criterion is expressed with squared loadings defined as correlation ratios for categorical variables and squared correlations for numerical variables. The key point of this rotation procedure is the definition of the single-plane rotation step. For rotation in more than two dimensions, the idea is to rotate pairs of dimensions (for which the optimal angle can be written) until the process converges. This “varimax type” rotation procedure is implemented in the function `PCArrot` package.

#### 4.1 The PCArrot algorithm

Suppose that  $k$  is the number of components required in a PCAmix interpretation. Let  $\mathbf{Z}$ ,  $\mathbf{N}$  and  $\mathbf{M}$  denote the matrices defined in the pre-processing step of the PCAmix procedure. The iterative rotation procedure gives the  $n \times k$  matrix  $\mathbf{F}_{\text{rot}}$  of the rotated factor scores of the rows of  $\mathbf{Z}$  and the  $(p_1 + m) \times k$  matrix  $\mathbf{A}_{\text{rot}}$  of the rotated factor scores of the columns of  $\mathbf{Z}$ . This procedure is carried out using the following steps.

**The initialization step.** The standard SVD (with metrics  $\mathbb{I}_n$  and  $\mathbb{I}_p$ ) of  $\tilde{\mathbf{Z}} = \mathbf{N}^{1/2}\mathbf{Z}\mathbf{M}^{1/2}$  gives the decomposition  $\tilde{\mathbf{Z}} = \tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{V}}^t$ . Therefore,  $\tilde{\mathbf{Z}} = \tilde{\mathbf{U}}\tilde{\mathbf{A}}^t$  where  $\tilde{\mathbf{A}} = \tilde{\mathbf{V}}\tilde{\mathbf{\Lambda}}$  denote the “loadings matrix” of this decomposition. Let us introduce  $\mathbf{T}$  an orthonormal rotation matrix. As  $\tilde{\mathbf{Z}} = \tilde{\mathbf{U}}\tilde{\mathbf{A}}^t = \tilde{\mathbf{U}}\mathbf{T}\mathbf{T}^t\tilde{\mathbf{A}}^t$ , the non-uniqueness of the solution of PCAmix can be exploited to improve the interpretability of the components.

The rotation procedure applies to the first  $k$  columns of  $\tilde{\mathbf{U}}$  and  $\tilde{\mathbf{A}}$ . To simplify the notations,  $\tilde{\mathbf{U}}$  and  $\tilde{\mathbf{A}}$  denote hereafter the truncated matrices of dimensions  $n \times k$  and  $(p_1 + m) \times k$ . Let  $\tilde{\mathbf{U}}_{\text{rot}}$  (resp.  $\tilde{\mathbf{A}}_{\text{rot}}$ ) denote the matrix  $\tilde{\mathbf{U}}$  (resp.  $\tilde{\mathbf{A}}$ ) after rotation. The initialization is  $\tilde{\mathbf{U}}_{\text{rot}} = \tilde{\mathbf{U}}$  and  $\tilde{\mathbf{A}}_{\text{rot}} = \tilde{\mathbf{A}}$ .

**The iterative optimization step.**

1. For each pair of dimensions  $(l, t)$  i.e. for  $l = 1, \dots, k - 1$  and  $t = (l + 1), \dots, k$ :

↪ calculate the angle of rotation  $\theta = \psi/4$  with:

$$\psi = \begin{cases} \arccos\left(\frac{h}{\sqrt{g^2 + h^2}}\right) & \text{if } g \geq 0, \\ -\arccos\left(\frac{h}{\sqrt{g^2 + h^2}}\right) & \text{if } g \leq 0, \end{cases} \quad (21)$$

where  $g$  and  $h$  are given by:

$$g = 2p \sum_{j=1}^p \alpha_j \beta_j - 2 \sum_{j=1}^p \alpha_j \sum_{j=1}^p \beta_j, \quad (22)$$

$$h = p \sum_{j=1}^p (\alpha_j^2 - \beta_j^2) - \left( \sum_{j=1}^p \alpha_j \right)^2 + \left( \sum_{j=1}^p \beta_j \right)^2, \quad (23)$$

with  $p$  the total number of variables, and  $\alpha_j$  and  $\beta_j$  defined by:

$$\alpha_j = \sum_{s \in I_j} (\tilde{a}_{sl, \text{rot}}^2 - \tilde{a}_{st, \text{rot}}^2) \quad \text{and} \quad \beta_j = 2 \sum_{s \in I_j} \tilde{a}_{sl, \text{rot}} \tilde{a}_{st, \text{rot}}. \quad (24)$$

Here,  $I_j$  is the set of row indices of  $\tilde{\mathbf{A}}_{\text{rot}}$  associated with the levels of the variable  $j$  in the categorical case and  $I_j = \{j\}$  in the numerical case.

↪ calculate the corresponding matrix of rotation  $\mathbf{T} = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix},$

↪ update the matrices  $\tilde{\mathbf{U}}_{\text{rot}}$  and  $\tilde{\mathbf{A}}_{\text{rot}}$  by rotation of their  $l$ -th and  $t$ -th columns.

2. Repeat the previous step until the  $k(k-1)/2$  successive rotations provide an angle of rotation  $\theta$  equal to zero.

### The rotated factor scores processing step.

1. The set of rotated factor scores for columns is computed as:

$$\mathbf{A}_{\text{rot}}^* = \mathbf{M}^{1/2} \tilde{\mathbf{A}}_{\text{rot}}. \quad (25)$$

$\mathbf{A}_{\text{rot}}^*$  is divided up as follows:  $\mathbf{A}_{\text{rot}}^* = \begin{bmatrix} \mathbf{A}_{1, \text{rot}}^* \\ \mathbf{A}_{2, \text{rot}}^* \end{bmatrix} \begin{matrix} \} p_1 \\ \} m \end{matrix}$  where

↪  $\mathbf{A}_{1, \text{rot}}^*$  contains the rotated factor scores of the  $p_1$  numerical variables,

↪  $\mathbf{A}_{2, \text{rot}}^*$  contains the rotated factor scores of the  $m$  levels.

2. The variance  $\lambda_{i, \text{rot}}$  of the  $i$ th rotated component is calculated as:

$$\lambda_{i, \text{rot}} = \|\tilde{\mathbf{a}}_{i, \text{rot}}\|_{\mathbb{P}}^2, \quad (26)$$

where  $\tilde{\mathbf{a}}_{i,\text{rot}}$  is the  $i$ th column of  $\tilde{\mathbf{A}}_{\text{rot}}$ . Let  $\Lambda_{\text{rot}} = \text{diag}(\sqrt{\lambda_{1,\text{rot}}}, \dots, \sqrt{\lambda_{k,\text{rot}}})$  denote the diagonal matrix of the standard deviations of the  $k$  rotated components. Note that after rotation, these eigenvalues will no longer necessarily appear in decreasing order.

3. The set of rotated factor scores for rows is computed as:

$$\mathbf{F}_{\text{rot}} = \mathbf{N}^{-1/2} \tilde{\mathbf{F}}_{\text{rot}}, \quad (27)$$

where  $\tilde{\mathbf{F}}_{\text{rot}} = \tilde{\mathbf{U}}_{\text{rot}} \Lambda_{\text{rot}}$ .

For numerical data, **PCAr<sub>rot</sub>** is the standard varimax procedure defined by Kaiser (1958) for rotation in PCA. For categorical data, **PCAr<sub>rot</sub>** is the corresponding “varimax type” rotation procedure for MCA.

## 4.2 Graphical outputs of PCAr<sub>rot</sub>

The properties used in PCAmix to interpret the component maps remain true after rotation:

- The rotated factor scores of the  $p_1$  numerical variables (the  $p_1$  first rows of  $\mathbf{A}_{\text{rot}}^*$ ) are correlations with the rotated principal components (the columns of  $\mathbf{F}_{\text{rot}}$ ).
- The rotated factor scores of the  $m$  categories (the  $m$  last rows of  $\mathbf{A}_{\text{rot}}^*$ ) are mean values of the (normalized) rotated factor scores of the observations that belong these levels.

The contribution of a variable  $j$  to the rotated component  $i$  can be calculated directly from the matrix  $\tilde{\mathbf{A}}_{\text{rot}}$  (instead of  $\mathbf{A}_{\text{rot}}^*$ ). We have:

$$\begin{cases} c_{ji,\text{rot}} = \tilde{a}_{ji,\text{rot}}^2 & \text{if the variable } j \text{ is numerical,} \\ c_{ji,\text{rot}} = \sum_{s \in I_j} \tilde{a}_{si,\text{rot}}^2 & \text{if the variable } j \text{ is categorical,} \end{cases} \quad (28)$$

If the variable  $j$  is numerical, we have  $c_{ji,\text{rot}} = r^2(\mathbf{f}_{i,\text{rot}}, \mathbf{x}_j)$  and if the variable  $j$  is categorical, we have  $c_{ji,\text{rot}} = \eta^2(\mathbf{f}_{i,\text{rot}} | \mathbf{x}_j)$ . In the context of rotation, the squared correlation and the correlation ratio are still called the “squared loadings”.

These properties are useful in interpreting the component maps of PCAmix after rotation.

## 4.3 Numerical outputs of PCAr<sub>rot</sub>

**PCAr<sub>rot</sub>** computes  $k$  new numerical variables called rotated principal components that “explain” the same part of the inertia of the data table  $\mathbf{Z}$  than PCAmix but with a simpler interpretation. Let us show that the rotated principal components (columns of  $\mathbf{F}_{\text{rot}}$ ) are still a (non-correlated) linear combination of the columns of  $\mathbf{Z}$  (new synthetic variables). First let us show that:

$$\mathbf{F}_{\text{rot}} = \mathbf{ZV}_{\text{rot}}, \quad (29)$$

with

$$\mathbf{V}_{\text{rot}} = \mathbf{M}^{1/2} \tilde{\mathbf{V}} \tilde{\mathbf{\Lambda}}^{-1} \mathbf{T} \mathbf{\Lambda}_{\text{rot}}, \quad (30)$$

and

$$\mathbf{T} = \tilde{\mathbf{U}}^t \tilde{\mathbf{U}}_{\text{rot}}. \quad (31)$$

**Proof:** The  $k \times k$  rotation matrix  $\mathbf{T}$  is such that

$$\tilde{\mathbf{U}}_{\text{rot}} = \tilde{\mathbf{U}} \mathbf{T}. \quad (32)$$

By definition of  $\tilde{\mathbf{U}}$ , we have  $\tilde{\mathbf{U}}^t \tilde{\mathbf{U}} = \mathbb{I}_r$ . It gives (31). By definition,  $\tilde{\mathbf{F}}_{\text{rot}} = \tilde{\mathbf{U}}_{\text{rot}} \mathbf{\Lambda}_{\text{rot}}$ . It gives  $\tilde{\mathbf{F}}_{\text{rot}} = \tilde{\mathbf{U}} \mathbf{T} \mathbf{\Lambda}_{\text{rot}}$ . The SVD decomposition  $\tilde{\mathbf{Z}} = \tilde{\mathbf{U}} \tilde{\mathbf{\Lambda}} \tilde{\mathbf{V}}^t$  writes also  $\tilde{\mathbf{U}} = \tilde{\mathbf{Z}} \tilde{\mathbf{V}} \tilde{\mathbf{\Lambda}}^{-1}$  (as  $\tilde{\mathbf{V}}^t \tilde{\mathbf{V}} = \mathbb{I}_r$  by definition). Then  $\tilde{\mathbf{F}}_{\text{rot}} = \tilde{\mathbf{Z}} \tilde{\mathbf{V}} \tilde{\mathbf{\Lambda}}^{-1} \mathbf{T} \mathbf{\Lambda}_{\text{rot}}$ . With  $\tilde{\mathbf{F}}_{\text{rot}} = \mathbf{N}^{1/2} \mathbf{F}_{\text{rot}}$  and  $\tilde{\mathbf{Z}} = \mathbf{N}^{1/2} \mathbf{Z} \mathbf{M}^{1/2}$ , it gives (29) and (30). ■

It follows that the linear combination of the vectors  $\mathbf{z}_1, \dots, \mathbf{z}_{p_1+m}$  (columns of  $\mathbf{Z}$ ) for the  $i$ th rotated principal component is:

$$\mathbf{f}_{i,\text{rot}} = \mathbf{Z} \mathbf{v}_{i,\text{rot}} = \sum_{k=1}^{p_1+m} v_{ki,\text{rot}} \mathbf{z}_k. \quad (33)$$

**Prediction of rotated scores of new observations.** The coefficients of these linear combinations are useful in projecting new observations onto the rotated principal components. It is easy to show that  $\mathbf{f}_{i,\text{rot}}$  writes:

$$\mathbf{f}_{i,\text{rot}} = \beta_{0,\text{rot}} + \sum_{k=1}^{p_1+m} \beta_{k,\text{rot}} \mathbf{x}_k, \quad (34)$$

where the vectors  $\mathbf{x}_1, \dots, \mathbf{x}_{p_1+m}$  are the columns of  $\mathbf{X} = (\mathbf{X}_1 | \mathbf{G})$ . It gives:

$$\begin{aligned} \beta_0 &= - \sum_{l=1}^{p_1} v_{li,\text{rot}} \frac{\bar{\mathbf{x}}_l}{\sigma_l} - \sum_{l=p_1+1}^{p_1+m} v_{li,\text{rot}} \frac{n}{n_l} \bar{\mathbf{x}}_l, \\ \beta_k &= v_{ki,\text{rot}} \frac{1}{\sigma_k}, \text{ for } k = 1, \dots, p_1, \\ \beta_k &= v_{ki,\text{rot}} \frac{n}{n_k}, \text{ for } k = p_1 + 1, \dots, p_1 + m, \end{aligned}$$

with  $\bar{\mathbf{x}}_k$  and  $\sigma_k$  respectively the empirical mean and standard deviation of the column  $\mathbf{x}_k$ .



#### 4.4 Illustration of PCArrot

The procedure PCArrot is applied to the same data `housing` as those used with PCAmix in Section 3.4. The three first principal components of PCAmix are rotated with the `PCArrot` function available in the R package PCAmixdata. Because the rotation modifies the interpretation of the third component, Fig. 2 plots the factor scores for the first and third components before and after rotation. After rotation Fig. 2(d) shows that the variable density becomes well projected and strongly associated to the third principal component. A new element of interpretation is then provided with rotation : a distinction of cities according to their density of population. Note that the benefits of using rotation on this dataset are quite limited. However we highlight the fact that the objective of this paper is not to demonstrate the practical advantages of rotation but to illustrate the use of the function `PCArrot`. The R code required to run this example is provided in Appendix D.

### 5 Multiple Factor Analysis of mixed data

Multiple Factor Analysis (MFA; Escofier and Pagès, 1994; Abdi et al., 2013) is used to analyze a set of observations described by several groups of variables. This method is implemented in the `MFA` function of the FactoMineR R package. The goal of MFA is to integrate different groups of variables describing the same observations. The straightforward analysis obtained by concatenating all variables would be dominated by the group with the strongest structure. The main idea of MFA is therefore to make all the groups of variables comparable in the analysis by introducing weights. The weight of a variable in the analysis will be the inverse of the variance of the first principal component of its group.

In standard MFA, the nature of the variables (categorical or numerical) can vary from one group to another, but the variables should be of the same type within a given group. The MFAmix method proposed here is able to handle mixed data within a group of variables.

#### 5.1 The MFAmix algorithm

Here the  $p$  variables are separated into  $G$  groups. The types of variables within a group can be mixed. Each group is represented by a data matrix  $\mathbf{X}^{(g)} = [\mathbf{X}_1^{(g)}, \mathbf{X}_2^{(g)}]$  where  $\mathbf{X}_1^{(g)}$  (resp.  $\mathbf{X}_2^{(g)}$ ) contains the numerical (resp. categorical) variables of group  $g = 1, \dots, G$ . The numerical columns (resp. the categorical columns) of the matrices  $\mathbf{X}^{(g)}$  are concatenated in a global numerical data matrix  $\mathbf{X}_1 = [\mathbf{X}_1^{(1)}, \dots, \mathbf{X}_1^{(G)}]$  (resp. a global categorical data matrix  $\mathbf{X}_2 = [\mathbf{X}_2^{(1)}, \dots, \mathbf{X}_2^{(G)}]$ ). Let  $\mathbf{Z}$ ,  $\mathbf{N}$  and  $\mathbf{M}$  denote the matrices constructed with  $\mathbf{X}_1$  and  $\mathbf{X}_2$  as described in the pre-processing

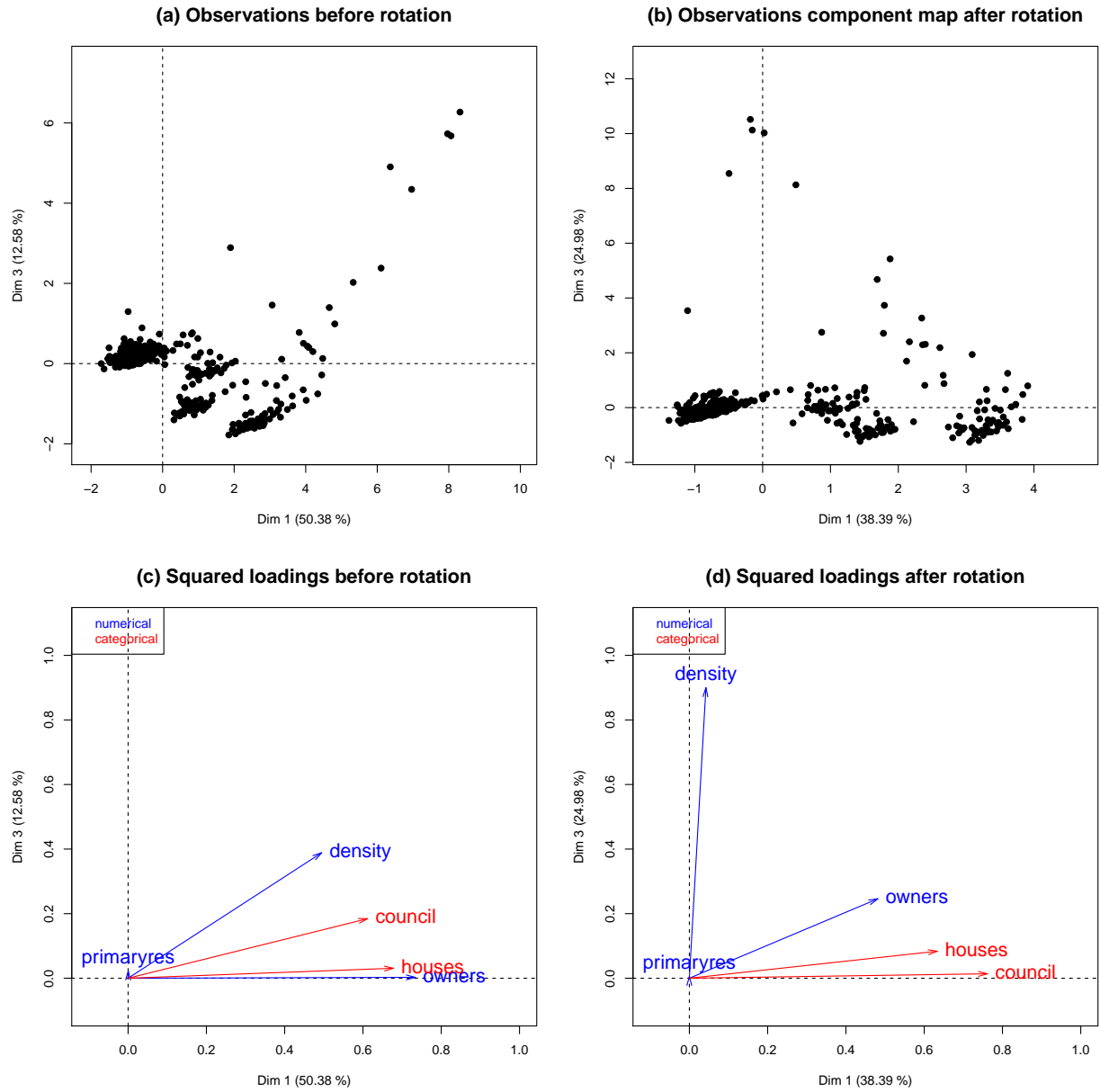


Figure 2: Graphical outputs of PCArot.

step of PCAmix. The MFAmix algorithm works as follows.

**Step 1 : weighting step.**

1. For  $g = 1, \dots, G$ , compute the first eigenvalue  $\lambda_1^{(g)}$  of PCAmix applied to  $\mathbf{X}^{(g)}$ .
2. Build the diagonal matrix  $\mathbf{P}$  of the weights of the columns of  $\mathbf{Z}$ . The weights are  $\frac{1}{\lambda_1^{(t_k)}}$ , where  $t_k \in \{1, \dots, g, \dots, G\}$  is the group of the  $k$ th column of  $\mathbf{Z}$ .

**Step 2: the factor scores processing step.**

1. The GSVD of  $\mathbf{Z}$  with metrics  $\mathbf{N}$  and  $\mathbf{M}^* = \mathbf{M}\mathbf{P}$  gives the decomposition:

$$\mathbf{Z} = \mathbf{U}_{\text{mfa}} \mathbf{\Lambda}_{\text{mfa}} \mathbf{V}_{\text{mfa}}^t$$

as defined in Subsection 2.1.

2. The set of factor scores for rows is computed as:

$$\mathbf{F}_{\text{mfa}} = \mathbf{U}_{\text{mfa}} \mathbf{\Lambda}_{\text{mfa}}. \quad (35)$$

3. The set of factor scores for columns is computed as:

$$\mathbf{A}_{\text{mfa}}^* = \mathbf{M}\mathbf{V}_{\text{mfa}} \mathbf{\Lambda}_{\text{mfa}}. \quad (36)$$

Note that the procedure for building the factor scores in MFAmix is different from that of PCAmix. Indeed, in MFAmix the matrix  $\mathbf{M}^* = \mathbf{M}\mathbf{P}$  is used in the GSVD whereas the matrix  $\mathbf{M}$  is used in PCAmix.

## 5.2 Graphical outputs of MFAmix

Naturally, MFAmix has the same properties as PCAmix when interpreting component maps. However the weights of the columns (the diagonal terms of  $\mathbf{M}^* = \mathbf{M}\mathbf{P}$ ) in MFAmix are different from those in PCAmix (the diagonal terms of  $\mathbf{M}$ ). The contribution of a variable  $j$  to a component  $i$  is now calculated from  $\mathbf{A}_{\text{mfa}} = \mathbf{V}_{\text{mfa}} \mathbf{\Lambda}_{\text{mfa}}$  as follows:

$$\begin{cases} c_{ji,\text{mfa}} = \frac{1}{\lambda_1^{(t_j)}} a_{ji,\text{mfa}}^2 = \frac{1}{\lambda_1^{(t_j)}} a_{ji,\text{mfa}}^{*2} & \text{if the variable } j \text{ is numerical,} \\ c_{ji,\text{mfa}} = \sum_{s \in I_j} \frac{1}{\lambda_1^{(t_s)}} \frac{n}{n_s} a_{si,\text{mfa}}^2 = \sum_{s \in I_j} \frac{1}{\lambda_1^{(t_s)}} \frac{n_s}{n} a_{si,\text{mfa}}^{*2} & \text{if the variable } j \text{ is categorical.} \end{cases} \quad (37)$$

Note that in MFAmix, the contribution of a categorical variable is no longer equal to the correlation ratio with the principal component, as previously in PCArrot and PCAmix.

**Contribution of a group.** The contribution of a group  $g$  is the sum of all the contributions of the variables of the group. The groups can then be plotted as points on a map using their contribution to the component coordinates.

**Partial observations.** The component map of the observations reveals the common structure through the groups but it is not possible to see how each group “interprets” the principal component space. The visualization of an observation according to a specific group (called a partial observation) can be achieved by projecting the dataset of each group onto this space. This is implemented in the following way.

1. For  $g = 1, \dots, G$ , construct the matrix  $\mathbf{Z}_{\text{part}}^{(g)}$  by putting to zero in  $\mathbf{Z}$  the values of the columns  $k$  such that  $t_k \neq g$ . The rows of  $\mathbf{Z}_{\text{part}}^{(g)}$  are the partial observations for the group  $g$ .
2. For  $g = 1, \dots, G$ , the factor scores for the partial observations are computed as:

$$\mathbf{F}_{\text{part}}^{(g)} = G \times \mathbf{Z}_{\text{part}}^{(g)} \mathbf{M}^* \mathbf{V}. \quad (38)$$

The scores of the partial observations are the coordinates of the orthogonal projections (with respect to the inner product matrix  $\mathbf{M}^*$ ) of the  $n$  rows of  $\mathbf{Z}_{\text{part}}^{(g)}$  onto the axes spanned by the columns of  $\mathbf{V}$  (with the the number of groups as multiplying factor).

The partial observations can then be plotted as supplementary points on the component map of the observations. The position of each observation is the barycenter of its positions for the partial observations. To facilitate this interpretation, lines linking an observation with the projection of its partial observations are drawn on the map.

**Partial axes.** MFAmix starts by applying PCAmix to the separated datasets  $\mathbf{X}^{(g)}$  for  $g = 1, \dots, G$ . Let  $\mathbf{f}_i^{(g)}$  denote the  $i$ th principal component of the analysis of  $\mathbf{X}^{(g)}$ . The axes of these separated PCAs are called the partial axes. Let  $\mathbf{f}_i$  denote the  $i$ th principal component of the global analysis (i.e. of MFAmix). The link between the separated analysis and the global analysis is explored by computing correlations between the principal components of each separated study and the principal components of the global study. The correlations  $r(\mathbf{f}_i^{(g)}, \mathbf{f}_i)$  are used as coordinates to plot the partial axes on a map.

### 5.3 Numerical outputs of MFAmix

The principal components of MFAmix are new numerical variables. The linear combination of the vectors  $\mathbf{z}_1, \dots, \mathbf{z}_{p_1+m}$  (columns of  $\mathbf{Z}$ ) for the  $i$ th principal component of MFAmix is:

$$\mathbf{f}_{i,\text{mfa}} = \mathbf{Z}\mathbf{M}^*\mathbf{v}_{i,\text{mfa}} = \sum_{k=1}^{p_1} \frac{1}{\lambda_1^{(t_k)}} v_{ki,\text{mfa}} \mathbf{z}_k + \sum_{k=p_1+1}^{p_1+m} \frac{1}{\lambda_1^{(t_k)}} \frac{n}{n_k} v_{ki,\text{mfa}} \mathbf{z}_k.$$

**Prediction of the scores of new observations** The coefficients of these linear combinations are useful in projecting new observations onto the principal components of MFAmix. It is easy to show that  $\mathbf{f}_i$  writes:

$$\mathbf{f}_{i,\text{mfa}} = \beta_0 + \sum_{k=1}^{p_1+m} \beta_k \mathbf{x}_k \quad (39)$$

where the vectors  $\mathbf{x}_1, \dots, \mathbf{x}_{p_1+m}$  are the columns of  $\mathbf{X} = (\mathbf{X}_1|\mathbf{G})$ . It gives:

$$\begin{aligned} \beta_0 &= -\sum_{l=1}^{p_1} v_{li,\text{mfa}} \frac{\bar{\mathbf{x}}_l}{\sigma_l} \frac{1}{\lambda_1^{(t_k)}} - \sum_{l=p_1+1}^{p_1+m} v_{li,\text{mfa}} \frac{1}{\lambda_1^{(t_k)}} \frac{n}{n_l} \bar{\mathbf{x}}_l, \\ \beta_k &= v_{ki,\text{mfa}} \frac{1}{\sigma_k} \frac{1}{\lambda_1^{(t_k)}}, \text{ for } k = 1, \dots, p_1, \\ \beta_k &= v_{ki,\text{mfa}} \frac{n}{n_k} \frac{1}{\lambda_1^{(t_k)}}, \text{ for } k = p_1 + 1, \dots, p_1 + m, \end{aligned}$$

where  $\bar{\mathbf{x}}_k$  and  $\sigma_k$  denote respectively the empirical mean and standard deviation of the column  $\mathbf{x}_k$ .

### 5.4 Illustration of MFAmix

The procedure MFAmix is illustrated using the `gironde` data where 542 cities are described with 27 variables separated into 4 groups (Employment, Housing, Services, Environment). The R code required to run this example is provided in Appendix E.

Fig. 3(a) shows the correlation circle of the global analysis with numerical variables colored by their group membership. Fig. 3(b) is a plot of the components of the separated PCAs where the coordinates are their correlations with the components of MFAmix. This plot suggests that the first principal component of the separated PCAmix on the group Services and Housing are highly correlated with the first principal component of MFAmix. On the other hand, only the first principal component of the group Employment is correlated with the second principal component of MFAmix. The first principal component of the group Environment is highly correlated with the third principal component of MFAmix (the graph is not presented for the sake of simplicity). Fig. 3(c) is a plot of the groups where a coordinate is the sum of the absolute contributions of the variables of the group. It suggests that the variables of the group Services and Housing both

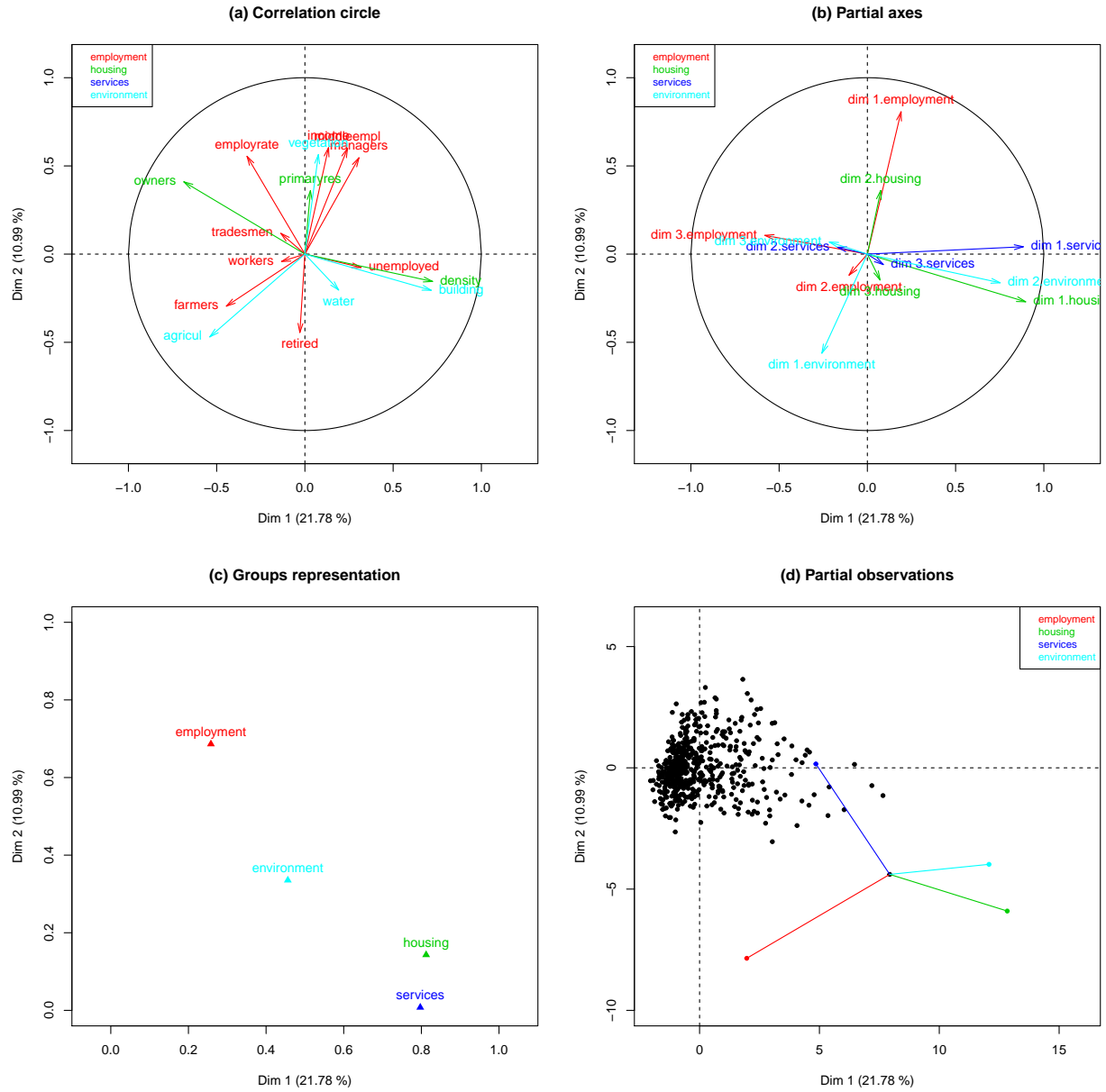


Figure 3: Graphical outputs of MFAMix

significantly contribute to the construction of the first principal component of MFAmix. More precisely, Fig. 3(a) shows that densely populated cities have (unsurprisingly) a high proportion of buildings. On the contrary we see that the variables “employrate”, “income” and “middleemp” are strongly correlated with the second principal component. This highlights the fact that cities with a high rate of employment are mostly composed of middle-range employees on average incomes. Fig. 3(d) represents the four partial observations (one for each group) of the city named Sainte-Foy-La-Grande. This city has an outsider position on the first factorial plane issued from MFAmix. This is mostly explained by its high coordinate on the first principal component. This suggests that this city has high group housing values, meaning that it has a dense population. When focusing on the group Services, this city has also a greater mean value than any other city within the Gironde region. This means that the number of services (bakers, dentists, doctors, restaurants, etc.) in this city is relatively high. However when considering the second principal component, this city has a low position, which is driven by the group Employment. This reveals that the rate of employment is low in this city, making it a place where people simply live, without benefiting from a prosperous job market.

## 6 Concluding remarks

This paper presented PCAmixdata, an R package designed to analyze a mixture of numerical and categorical variables. Three methods are implemented for the analysis of such data in the functions `PCAmix`, `PCarot` and `MFAmix`. We provided a synthetic presentation of the technical details of these approaches, and described the numerical and graphical outputs of the package. These results were illustrated using the real dataset `gironde` describing socio-economic and environmental characteristics of cities in a region of south-west of France.

Some extensions to the package are currently under study. A first simple extension will be the possibility of using rotation in MFAmix. It would also be interesting to implement discriminant analysis functions for mixtures of numerical and categorical variables.

The Comprehensive R Archive Network <http://cran.r-project.org/> provides the package, including sources, binaries and documentation, for download under the GNU Public License.

# Appendices

## A Proof of equation (13)

We demonstrate here how to obtain MCA from a single PCA with metrics. MCA is a CA applied to the indicator matrix  $\mathbf{G}$ . Let  $\mathbf{F}$  denote the matrix of the factor scores of the observations and  $\mathbf{A}^*$  denote the matrix of the factor scores of the categories. The factor scores of the observations and the factor scores of the categories are obtained by applying two different PCAs, one applied to the row profiles and one applied to the column profiles. The row and columns profiles are calculated from the so-called ‘frequency’ matrix  $\frac{\mathbf{G}}{np}$ . Margins of this matrix are used to weight rows and columns in these two analysis. Let us introduce the following notations.

- $\mathbf{r} \in \mathbb{R}^n$  is the vector of the weights of the observations. The weight of a row is constant and is equal to  $\frac{1}{n}$ .
- $\mathbf{c} \in \mathbb{R}^m$  is the vector of the weights of the categories. The weight of a column is equal to  $\frac{n_s}{np}$  and grows for frequent categories.
- $\mathbf{D}_r = \text{diag}(\mathbf{r}) = \frac{1}{n}\mathbb{I}_n$  is the diagonal matrix of the weights of the observations.
- $\mathbf{D}_c = \text{diag}(\mathbf{c}) = \text{diag}(\frac{n_s}{np}, s = 1, \dots, m)$  is the diagonal matrix of the weights of the categories.
- $\mathbf{L} = \mathbf{D}_r^{-1} \frac{\mathbf{G}}{np} = \frac{\mathbf{G}}{p}$  is the matrix of the profiles of the observations.
- $\mathbf{C} = \frac{\mathbf{G}}{np} \mathbf{D}_c^{-1}$  is the matrix of the profiles of the categories.

**PCA of the rows of  $\mathbf{L}$ .** The GSVD of  $\mathbf{L}$  with metrics  $\mathbf{D}_r$  and  $\mathbf{D}_c^{-1}$  gives the decomposition:

$$\mathbf{L} = \mathbf{U}_L \mathbf{\Lambda} \mathbf{V}_L^t, \quad (40)$$

where  $\mathbf{V}$  is the eigenvectors matrix of  $\mathbf{L}^t \mathbf{D}_r \mathbf{L} \mathbf{D}_c^{-1}$  and  $\mathbf{U}_L$  is the matrix of eigenvectors of  $\mathbf{L} \mathbf{D}_c^{-1} \mathbf{L}^t \mathbf{D}_r$ . It gives:

$$\begin{aligned} \mathbf{F} &= \mathbf{L} \mathbf{D}_c^{-1} \mathbf{V}_L = \mathbf{U}_L \mathbf{\Lambda}, \\ \mathbf{A}_L &= \mathbf{L}^t \mathbf{D}_r \mathbf{U}_L = \mathbf{V}_L \mathbf{\Lambda}, \end{aligned}$$

where  $\mathbf{F}$  denotes the factor scores of the rows of  $\mathbf{L}$  (the row profiles) and  $\mathbf{A}_L$  denotes the factor scores of the columns of  $\mathbf{L}$ . Note that the columns of  $\mathbf{L}$  are not the column profiles.



**PCA of the columns of  $\mathbf{C}$ .** The GSVD of  $\mathbf{C}$  with metrics  $\mathbf{D}_r^{-1}$  and  $\mathbf{D}_c$  gives the decomposition:

$$\mathbf{C} = \mathbf{U}_C \mathbf{\Lambda} \mathbf{V}_C^t, \quad (41)$$

where  $\mathbf{U}_C$  is the eigenvectors matrix of  $\mathbf{C} \mathbf{D}_c \mathbf{C}^t \mathbf{D}_r^{-1}$  and  $\mathbf{V}_C$  is the eigenvectors matrix of  $\mathbf{C}^t \mathbf{D}_r^{-1} \mathbf{C} \mathbf{D}_c$ . It gives:

$$\begin{aligned} \mathbf{A}^* &= \mathbf{C}^t \mathbf{D}_r^{-1} \mathbf{U}_C = \mathbf{V}_C \mathbf{\Lambda}, \\ \mathbf{F}_C &= \mathbf{C} \mathbf{D}_c \mathbf{V}_C = \mathbf{U}_C \mathbf{\Lambda}, \end{aligned}$$

where  $\mathbf{A}^*$  denotes the factor scores of the columns of  $\mathbf{C}$  (the column profiles) and  $\mathbf{F}_C$  denotes the factor scores of the rows of  $\mathbf{C}$  (which are not the row profiles).

**A single PCA of the indicator matrix  $\mathbf{G}$ .** At this level, the matrices  $\mathbf{F}$  and  $\mathbf{A}^*$  are calculated from two different PCAs. We prove below that  $\mathbf{V}_C = \mathbf{D}_c^{-1} \mathbf{V}_L$ . It follows that  $\mathbf{A}^* = \mathbf{D}_c^{-1} \mathbf{V}_L \mathbf{\Lambda}$  is obtained directly from the GSVD of  $\mathbf{L}$ .

**Proof.** Let us demonstrate now that  $\mathbf{V}_C = \mathbf{D}_c^{-1} \mathbf{V}_L$ . Matrices  $\mathbf{L}^t \mathbf{D}_r \mathbf{L} \mathbf{D}_c^{-1}$  and  $\mathbf{C}^t \mathbf{D}_r^{-1} \mathbf{C} \mathbf{D}_c$  have the same eigenvalues. Let  $\mathbf{\Lambda}$  denote the diagonal matrix of these eigenvalues.  $\mathbf{V}_L$  is the eigenvectors matrix of  $\mathbf{L}^t \mathbf{D}_r \mathbf{L} \mathbf{D}_c^{-1}$ , then we have:

$$(\mathbf{L}^t \mathbf{D}_r \mathbf{L} \mathbf{D}_c^{-1}) \mathbf{V}_L = \mathbf{\Lambda} \mathbf{V}_L. \quad (42)$$

$\mathbf{V}_C$  is the eigenvector matrix of  $\mathbf{C}^t \mathbf{D}_r^{-1} \mathbf{C} \mathbf{D}_c$ , then we have:

$$(\mathbf{C}^t \mathbf{D}_r^{-1} \mathbf{C} \mathbf{D}_c) \mathbf{V}_C = \mathbf{\Lambda} \mathbf{V}_C. \quad (43)$$

$\mathbf{L} = \mathbf{D}_r^{-1} \mathbf{C} \mathbf{D}_c$ . From that expression of  $\mathbf{L}$ , we can write  $\mathbf{L}^t \mathbf{D}_r \mathbf{L} \mathbf{D}_c^{-1}$  according to  $\mathbf{C}$ :

$$\begin{aligned} \mathbf{L}^t \mathbf{D}_r \mathbf{L} \mathbf{D}_c^{-1} &= (\mathbf{D}_r^{-1} \mathbf{C} \mathbf{D}_c)^t \mathbf{D}_r (\mathbf{D}_r^{-1} \mathbf{C} \mathbf{D}_c) \mathbf{D}_c^{-1} \\ &= \mathbf{D}_c \mathbf{C}^t \mathbf{D}_r^{-1} \mathbf{D}_r \mathbf{D}_r^{-1} \mathbf{C} \mathbf{D}_c \mathbf{D}_c^{-1} \\ &= \mathbf{D}_c \mathbf{C}^t \mathbf{D}_r^{-1} \mathbf{C}. \end{aligned} \quad (44)$$

$\mathbf{B} = \mathbf{L}^t \mathbf{D}_r \mathbf{L} \mathbf{D}_c^{-1}$  is  $\mathbf{D}_c^{-1}$  symmetric, then we have:

$$\mathbf{D}_c^{-1} \mathbf{B} = \mathbf{B}^t \mathbf{D}_c^{-1}. \quad (45)$$

From (42) we obtain  $\mathbf{B} \mathbf{V}_L = \mathbf{\Lambda} \mathbf{V}_L$  and then  $\mathbf{D}_c^{-1} \mathbf{B} \mathbf{V}_L = \mathbf{\Lambda} \mathbf{D}_c^{-1} \mathbf{V}_L$ .

From (45), we obtain  $\mathbf{B}^t \mathbf{D}_c^{-1} \mathbf{V}_L = \mathbf{\Lambda} \mathbf{D}_c^{-1} \mathbf{V}_L$ .

By using (44) to rewrite  $\mathbf{B}$ , we have:

$$(\mathbf{D}_c \mathbf{C}^t \mathbf{D}_r^{-1} \mathbf{C})^t \mathbf{D}_c^{-1} \mathbf{V}_L = \Lambda \mathbf{D}_c^{-1} \mathbf{V}_L$$

and  $(\mathbf{C}^t \mathbf{D}_r^{-1} \mathbf{C} \mathbf{D}_c) \mathbf{D}_c^{-1} \mathbf{V}_L = \Lambda \mathbf{D}_c^{-1} \mathbf{V}_L$ .

By identification in (43), we obtain  $\mathbf{V}_C = \mathbf{D}_c^{-1} \mathbf{V}_L$ .

■

## B Variables of the data gironde

## C R code for the PCAmix procedure

The following code runs the example described in Section 3.4. The illustration is made with the datatable `housing` of the dataset `gironde` available in the package `PCAmixdata`. The ten first rows (cities) are not used in first stage in `PCAmix` in order to predict later their factor scores with the function `predictPCAmix`.

```
#Load the package and the data
library(PCAmixdata)
data(gironde)

#Create the datatable housing without the ten first observations
housing<-gironde$housing[-c(1:10), ]

#Split the datatable
split<-splitmix(housing) # numerical data
X1<-split$X.quanti; X2<-split$X.quali #categorical data

#Perform PCAmix.
res.pcamix<-PCAmix(X.quanti=X1, X.quali=X2, rename.level=TRUE, graph=FALSE)

#Numerical output of PCAmix
res.pcamix$eig #eigenvalues, percentages of variance
res.pcamix$ind$coord #coordinates of observations
res.pcamix$levels$coord #coordinates of levels
res.pcamix$quali$contrib.pct #relative contributions of categorical variables
res.pcamix$coef #coefficients of the linear combinations defining the PCs

#Prediction of the coordinates of the 10 first cities
```

<b>R_Names</b>	<b>Description</b>	<b>Group</b>	<b>Data type</b>
farmers	Percentage of farmers	employment	Num
tradesmen	Percentage of tradesmen and shopkeepers	employment	Num
managers	Percentage of managers and executives	employment	Num
workers	Percentage of workers and employees	employment	Num
unemployed	Percentage of unemployed workers	employment	Num
middleemp	Percentage of middle-range employees	employment	Num
retired	Percentage of retired people	employment	Num
employrate	employment rate	employment	Num
income	Average income	employment	Num
density	Population density	housing	Num
primaryres	Percentage of primary residences	housing	Num
houses	Percentage of houses	housing	Categ
owners	Percentage of home owners living in their primary residence	housing	Num
council	Percentage of council housing	housing	Categ
butcher	Number of butchers	services	Categ
baker	Number of bakers	services	Categ
postoffice	Number of post offices	services	Categ
dentist	Number of dentists	services	Categ
grocery	Number of grocery stores	services	Categ
nursery	Number of child care day nurseries	services	Categ
doctor	Number of doctors	services	Categ
chemist	Number of chemists	services	Categ
restaurant	Number of restaurants	services	Categ
building	Percentage of buildings	environment	Num
water	Percentage of water	environment	Num
vegetation	Percentage of vegetation	environment	Num
agricul	Percentage of agricultural land	environment	Num

```

newind<-gironde$housing[1:10, ]
splitnew<-splitmix(newind)
X1new<-splitnew$X.quanti; X2new<-splitnew$X.quali
pred<-predict.PCAmix(object=res.pcamix, X.quanti=X1new, X.quali=X2new)

#Graphical output of PCAmix
houses <- X2$houses
plot(res.pcamix,choice="ind",axes=c(1,2),coloring.ind=houses,label=FALSE,
      posleg="bottomright", main="Observations")
plot(res.pcamix,choice="levels",axes=c(1,2),xlim=c(-1.5,2.5),cex=0.9, main="Levels")
plot(res.pcamix,choice="cor",axes=c(1,2),main="Numerical variables")
plot(res.pcamix,choice="sqload",axes=c(1,2), coloring.var="type", leg=TRUE, xlim=c(-0.1,1.05),
      posleg="topright", main="All variables")

```

## D R code for the PCArrot procedure

The following code runs the example described in Section 4.4.

```

#Load the package and the data
library(PCAmixdata)
data(gironde)

#Create the datatable housing without the ten first observations
housing<-gironde$housing[-c(1:10), ]
#Split the datatable
split<-splitmix(housing) # numerical data
X1<-split$X.quanti; X2<-split$X.quali #categorical data

#Perform PCAmix.
res.pcamix<-PCAmix(X.quanti=X1, X.quali=X2, rename.level=TRUE, graph=FALSE)

#Orthogonal rotation of the three first principal component
res.pcarot<-PCArrot(res.pcamix,dim=3,graph=FALSE)

#Numerical output of PCArrot
res.pcarot$eig #variance of the rotated PCs
res.pcarot$ind$coord #coordinates of observations on the rotated components

```

```

res.pcarot$levels$coord #coordinates of numerical variables on the rotated components
res.pcarot$coef #coefficients of the linear combinations defining the rotated PCs

#Prediction of the coordinates of the 10 first cities on the rotated components
newind<-gironde$housing[1:10, ]
splitnew<-splitmix(newind)
X1new<-splitnew$X.quanti; X2new<-splitnew$X.quali
pred.rot<-predict.PCAmix(object=res.pcarot, X.quanti=X1new, X.quali=X2new)

#Graphical output of PCArrot
plot(res.pcarot,choice="ind",label=FALSE,axes=c(1,3),
      main="Observations after rotation")
plot(res.pcarot,choice="sqload", coloring.var="type", leg=TRUE,axes=c(1,3),
      posleg="topright", main="Squared loadings after rotation")

```

## E R code for the MFAmix procedure

The following code runs the example described in section 5.4.

```

#Load the package and the data
library(PCAmixdata)
data(gironde)

#Concatenation of the 4 datatables
dat<-cbind(gironde$employment,gironde$housing,gironde$services,gironde$environment)
class.var<-c(rep(1,9),rep(2,5),rep(3,9),rep(4,4)) #definition of the groups of variables
names<-c("employment","housing","services","environment") #names of the groups of variables

#Perform MFAmix
res.mfamix<-MFAmix(data=dat,groups=class.var,
                   name.groups=names,ndim=3,rename.level=TRUE,graph=FALSE)

#Numerical output of MFAmix (among others)
res.mfamix$eig #eigenvalues, percentages of variance of the global analysis
res.mfamix$eig.separate #eigenvalues of the 4 separated PCAmix analysis
res.mfamix$group$coord #Coordinates of groups

#Prediction of the coordinates of the 10 first cities

```

```

newind<-dat[1:10, ]
pred.mfamix<-predict.MFAMix(object=res.mfamix, data=newind, groups=class.var, name.groups=names)

#Graphical output of MFAMix
plot(res.mfamix, choice="cor", coloring.var="groups", leg=TRUE,
      main="Correlation circle")
plot(res.mfamix, choice="axes", coloring.var="groups", leg=TRUE,
      main="Partial axes")
plot(res.mfamix, choice="groups", coloring.var="groups",
      main="Groups representation")
plot(res.mfamix, choice="ind", partial=c("SAINTE-FOY-LA-GRANDE"), label=FALSE,
      posleg="topright", main="Partial observations")

```

## References

- Abdi, H., Williams, L. J., and Valentin, D. (2013). Multiple factor analysis: principal component analysis for multitable and multiblock data sets. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(2):149–179.
- Beaton, D., Chin Fatt, C. R., and Abdi, H. (2014). An exposition of multivariate analysis with the singular value decomposition in r. *Computational Statistics & Data Analysis*, 72:176–189.
- Chavent, M., Kuentz-Simonet, V., and Saracco, J. (2012). Orthogonal rotation in pcamix. *Advances in Data Analysis and Classification*, 6(2):131–146.
- Dray, S., Dufour, A.-B., et al. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of statistical software*, 22(4):1–20.
- Escofier, B. and Pagès, J. (1994). Multiple factor analysis (afmult package). *Computational statistics & data analysis*, 18(1):121–140.
- Hill, M. and Smith, A. (1976). Principal component analysis of taxonomic data with multi-state discrete characters. *Taxon*, 25(2/3):249–255.
- Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200.
- Kiers, H. A. (1991). Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. *Psychometrika*, 56(2):197–212.

- Lê, S., Josse, J., and Husson, F. (2008). Factominer: an r package for multivariate analysis. *Journal of statistical software*, 25(1):1–18.
- Pagès, J. (2004). Analyse factorielle de données mixtes. *Revue de statistique appliquée*, 52(4):93–111.